WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

A HYBRID CLUSTERING ALGORITHM FOR IDENTIFYING HIGH DENSITY CLUSTERS

M. Anthony Wong

W.P. No. 1101-80 January, 1980

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139

A HYBRID CLUSTERING ALGORITHM FOR IDENTIFYING HIGH DENSITY CLUSTERS

M. Anthony Wong

W.P. No. 1101-80 January, 1980

ABSTRACT

High density clusters are defined on a population with density f to be the maximal connected sets of values x with $f(x) \geq c$, for various values of c. It is desired to discover the high density clusters given a random sample of size N from the population. Using this clustering model, there is a correspondence between clustering and density estimation techniques. A hybrid algorithm is proposed which combines elements of both the k-means and single linkage techniques. This procedure is practicable for very large number of observations, and is shown to be consistent, under certain regularity conditions in one dimension.

KEY WORDS: High density clusters; K-Means clustering; Single linkage; Density estimation; Asymptotic consistency; Hybrid clustering.

1. INTRODUCTION

The high density model for clusters (Hartigan 1975, p.205) assumes that observations $x_1, x_2, \ldots, x_N$ are sampled from a population with density f in p dimensional space, taken with respect to Lebesgue measure. High density clusters are maximal connected sets of the form $\{x \mid f(x) \geq c\}$, taken for all c. The family T of such clusters forms a tree, in that $A \in T$, $B \in T$ implies $A \supset B$, $B \supset A$ or $A \cap B = \phi$. A sample hierarchical clustering T_N on $x_1, x_2, \ldots, x_N$ may now be evaluated by how well it approximates T, on the average. It may be asked whether or not the sample clusters converge to the population clusters in some sense. The procedure T_N is <u>set-consistent</u> for T if for each $A, B \in T$ with $A \cap B = \phi$, there exists $A_N, B_N \in T_N$ with $A_N \supset A \cap \{x_1, \ldots, x_N\}$, $B_N \supset B \cap \{x_1, \ldots, x_N\}$, $A_N \cap B_N = \phi$, with probability approaching 1 as $N \to \infty$.

Standard hierarchical techniques begin with clusters consisting of single points, and successively join pairs of clusters which are closest according to some measure of distance, to obtain new clusters. The process terminates when a single cluster remains. Complete linkage (Sorenson 1948) defines distance between clusters to be the maximum distance of pairs of points in the clusters, and it is not set-consistent (Hartigan 1977a). Average linkage (Sneath and Sokal 1973) uses distance between clusters as the average distance between pairs of points in the two clusters, and sampling experiments (Hartigan 1979) suggest it is not

consistent either. Single linkage (Sneath 1957) defines distance between clusters as the closest distance between pairs of points in the two clusters. Single linkage is set consistent in one dimension but not in higher dimensions. There is empirical evidence and some theory (Hartigan 1979) to suggest that single linkage is consistent in a weaker sense.

Density estimates generate clusters, namely the high density clusters corresponding to the estimates. Single linkage corresponds to nearest neighbour density estimation (Hartigan 1977b), in which the density estimate at a point x is inversely proportional to the volume of the smallest closed sphere including a sample point. This density estimate is inconsistent in the sense that $f_N(x)$ does not approach $f(x)$ in probability. An improved density estimate, and perhaps improved clustering, is obtained by the kth nearest neighbour density estimate: the density at point x is inversely proportional to the volume of the smallest sphere containing k sample points. Such a density estimate is consistent at a point x if f is continuous at x and $k \to \infty$ as $N \to \infty$. More generally, kernel estimates of the form $\frac{1}{N}\Sigma_1^N K_N(x, x_i)$ might be used (see, for example, Wegman 1972).

Although the statistical justification of these density estimates require N very large, the number of computations is usually $O(N^2)$ which begins to be onerous for N over 250. In addition, the actual computation of high density cluster from the density estimate may be formidable. A variation of kth nearest neighbour from which clusters may be constructed is due to Wishart (1969); the density at observation x_j is kth nearest

neighbour density, points x_i and x_j are connected if x_j is among the k closest points to x_i, or if x_i is among the k closest points to x_j; and the high density sets with this measure of connectedness are clusters. The computational expense of this technique is $O(N^2)$. For related techniques, see Ling (1973) and Jardine and Sibson (1971).

A hybrid clustering technique is proposed here which combines the k-means (Hartigan and Wong 1979) and single linkage clustering techniques. At the first stage, k-means is used to construct a variable cell histogram (with k cells) which provides uniformly consistent estimates of the underlying density (Wong, 1980). At the second stage, the high density clusters corresponding to the computed estimates are obtained by applying single linkage to an appropriate distance matrix defined on the k cells. A detailed description of this method is given in Section 2. The number of calculations is O(Nk). In Section 3, it is shown that the hybrid algorithm is set-consistent for high density clusters in one dimension, under certain regularity conditions. Some empirical evidence are given in Section 4 to show that hybrid clustering is a useful tool for identifying high density clusters.

2. THE HYBRID ALGORITHM

The algorithm consists of two stages: At the first stage, the observations are clumped into k clusters by k-means, so that no movement of an observation from one cluster to another reduces the within cluster sum of squares. A histogram estimate is then constructed on the k regions defined by the k-means partition. At the second stage, the distance between neighbouring clusters is taken inversely proportional to the density estimate at a point halfway between the cluster means, and single linkage is applied to the distance matrix to obtain the tree of high density clusters corresponding to the histogram estimate of the density. In one dimension, the algorithm works as follows: a histogram consisting of k intervals in constructed on the k clusters obtained by k-means; in the second stage, using the computed density estimates, neighbouring clusters are then joined successively to give the tree of sample clusters. Since the k-means procedure provides a practicable and convenient way of obtaining a k-partition of multivariate data, the generalization to p dimensions ($p > 1$) is immediate.

2.1 The K-means step

A k-means partition will be taken to be a partition into k clusters such that no observations can be moved from one cluster to another without increasing the within cluster sum of squares. There are a number of ways of reaching such a partition by transferring observations to reduce within cluster sum of squares (see, for example, Hartigan and Wong 1979); the

number of computations is usually proportional to NkIp where N is the number of observations, k is the number of clusters, I is the number of iterations reallocating all observations, and p is the number of dimensions. The asymptotic properties of k-means as a clustering technique (as N approaches ∞ with k fixed) have been studied by MacQueen (1967), Hartigan (1978), and Pollard (1979). In this application, however, it is used primarily as a density estimation procedure.

The asymptotic properties of k-means as a procedure for providing a histogram estimate of the density are given in Wong (1980). In one dimension, the following density estimate is shown to be uniformly consistent in probability:

Lemma (Wong 1980, Corollary 7):

Let $x_1, \ldots, x_N$ be a random sample from some population F on $[a,b]$. Suppose that the density f is four times differentiable and is strictly positive on $[a, b]$. Consider a locally optimal k-means partition of the sample with k_N clusters. Let n_j be the number of observations in the jth cluster $(j = 1, \ldots, k_N)$.

And let $\bar{x}_j$ and WSS_j respectively be the sample mean and within cluster sum of squares of the jth cluster $(j= 1, \ldots, k_N)$.

Define the pooled density estimate at a point x between neigbouring cluster means $\bar{x}_{j-1}$ and $\bar{x}_j$ by :

$$f_N(x) = (n_j + n_{j-1})^{3/2} / N\,(12WSS_j^*)^{\frac{1}{2}},\ \bar{x}_{j-1} < x \leq \bar{x}_j \quad (j=2,\ldots, k_N);$$

$$= (n_2 + n_1)^{3/2} / N\,(12WSS_2^*)^{\frac{1}{2}},\ a \leq x \leq \bar{x}_1 ;$$

$$= (n_{k_N} + n_{k_N-1})^{3/2} / N\,(12WSS_{k_N}^*)^{\frac{1}{2}},\ \bar{x}_{k_N} < x \leq b,$$

where $WSS_j^* = WSS_j + WSS_{j-1} + \frac{1}{4}(n_j + n_{j-1}) \cdot (\bar{x}_j - \bar{x}_{j-1})^2$.

Then provided that $k_N = o\left([N/\log N]^{1/3} \right)$,

$$\sup_{a \leq x \leq b} |f_N(x) - f(x)| = o_p(1).$$

Unfortunately, the univariate results cannot be easily generalized to the multivariate case. However, let us assume that in many dimensions (R^p, $p > 1$), the ith cluster consists of a regular isotope of volume v_i centred on the cluster mean $\bar{x}_i$. Then $WSS_i \propto n_i v_i^{2/p}$ and $n_i \propto f(\bar{x}_i) v_i$. It follows that $f(\bar{x}_i) \propto n_i^{1+p/2} WSS_i^{-p/2}$, and hence $n_i^{1+p/2} WSS_i^{-p/2}$ can be interpreted as an estimate of $cf(\bar{x}_i)$ where c is some proportionality constant. And for adjoining clusters i and j, it is conjectured that a consistent pooled estimate of the density at x_{ij}, the midpoint between $\bar{x}_i$ and $\bar{x}_j$, is given by

$$f_N(x_{ij}) \propto (n_i + n_j)^{1+p/2} / \left[WSS_i + WSS_j + \tfrac{1}{4}(n_i + n_j) \cdot d^2(\bar{x}_i, \bar{x}_j) \right]^{p/2}, \quad (2.1)$$

where d is the Euclidean distance. (Note that when $p = 1$, $f_N(x_{ij})$ is the estimate given in the Lemma.) The assumptions are plausible in two dimensions as k-mean clusters are likely to be regular hexagons when k is large, but in three or more dimensions, it is not clear that the best partition is into regular isotopes. Here, the within cluster sums of squares are being used to measure the volume of the clusters, which is acceptable if all clusters are approximately the same shape. The volumes could be computed directly but at great computational expense in many dimensions. Much work has yet to be done to prove the conjecture for two or more dimensions.

2.2. The Single linkage step

The k-means step produces k clusters with cluster means $\bar{x}_1, \ldots, \bar{x}_k$. The single linkage step constructs hierarchical high density clusters on the k clusters using the density estimates f_N obtained in the k-means step. The following property of single linkage recommends its use in this cluster-formation stage of the algorithm: At a given distance level D*, any two objects in the same single linkage cluster can be connected by a chain of links of objects such that the size of each link is no greater than D*. Thus, if the distances between connected clusters are reciprocal to the density estimates f_N, every resulting single linkage cluster corresponds to a maximal connected region of the form $\{ x \mid f_N(x) \geq c \}$.

Hence, a distance matrix is computed for the k clusters as follows: Two clusters i and j are said to be connected if x_{ij}, the midpoint between $\bar{x}_i$ and $\bar{x}_j$, is closer to $\bar{x}_i$ (or $\bar{x}_j$) than any other cluster mean. If clusters i and j are connected, then $D(i,j) = f_N^{-1}(x_{ij})$; otherwise $D(i,j) = \infty$. (See (2.1) for definition of f_N). Single linkage clusters are then computed from this distance matrix to give the sample high density clusters.

Next, we will examine the asymptotic consistency of the hybrid algorithm for high density clusters.

3. CONSISTENCY OF HYBRID CLUSTERING FOR HIGH DENSITY CLUSTERS

Let f denote a density on [a,b] such that $\{ x \mid f(x) > c \}$ is the union of a finite number of closed intervals for every $c > 0$. Let T be the tree of population high density clusters defined on f. Let $x_1, \ldots, x_N$ be a random sample from f and let T_N be the hierarchical clustering specified by the hybrid algorithm.

Theorem: Suppose that A and B are any two disjoint high density clusters in T. Assume that f is positive and has four bounded derivatives in [a,b]. Then provided that $k_N = o\left([N/\log N]^{1/3} \right)$, there exist A_N, $B_N \in T_N$ with $A_N \supseteq A \cap \{X_1, \ldots, X_N\}$, $B_N \supseteq B \cap \{X_1, \ldots, X_N\}$, and $A_N \cap B_N = \phi$, with probability tending to 1 as $N \to \infty$.

Proof: Since T_N is the tree of high density clusters for f_N (see Lemma), this theorem is a direct consequence of the Lemma, which states that

$$\sup_{a \le x \le b} |f(x) - f_N(x)| = o_p(1). \tag{3.1}$$

By definition, for any two disjoint high density clusters A and B in T, there exist $\varepsilon > 0$ and $\lambda > 0$ such that

$$f(x) \ge \lambda \quad \text{for all } x \in A \cup B, \tag{3.2}$$

and A and B are separated by a region V, where

$$f(x) < \lambda - 3\varepsilon \quad \text{for all } x \in V. \tag{3.3}$$

From (3.1), we have

$$P_r \{ \sup_{a \le x \le b} | f(x) - f_N(x) | < \varepsilon \} \to 1 \ .$$

Thus, it follows from (3.2) and (3.3) that for N large, with high probability,

$$f_N(x) > \lambda - \varepsilon \qquad \text{for all } x \in A \cup B, \tag{3.4}$$

and $$f_N(x) < \lambda - 2\varepsilon \qquad \text{for all } x \in V. \tag{3.5}$$

Since A and B are disjoint, it follows from (3.4) and (3.5) that high density clusters of the form $\{x \mid f_N(x) \ge \lambda - \varepsilon\}$ separate A and B. The theorem follows.

The above theorem shows that the hybrid algorithm is set-consistent in one dimension, for densities f on $[a,b]$ which are positive and have four derivatives, and for k-means partitions into k_N clusters where $k_N^3 \log N/N \to \infty$, $k_N \to \infty$, as $N \to \infty$. We conjecture a similar result will hold in two dimensions. The higher dimensional case requires further study; empirical results suggest that hybrid algorithm is useful for identifying high density clusters.

4. EMPIRICAL VALIDATION OF THE HYBRID ALGORITHM

The hybrid method was applied to various generated data sets to test for its effectiveness in specifying high density clusters (Wong 1979). Results of three of the experiments, one using univariate data and the other two bivariate, are reported here.

1. Experiment One: In this experiment, 1000 observations are drawn from the univariate normal mixture $\frac{1}{2}N(0,1) + \frac{1}{2}N(3,1)$. This data set is useful in showing the performance of the hybrid algorithm when two high density clusters are separated by a region of moderate density. The density estimates over the intervals between the k=40 cluster means are plotted in Figure A. (A rough rule of thumb for k is $7(N/\log N)^{1/3}$.). Although the minimum density between the modes is more than half the density at the modes, the hybrid algorithm would still produce a hierarchical clustering which clearly indicates the presence of two modal regions (see Figure B).

2. Experiment Two: Here, a sample of size 1000 is taken from the bivariate normal mixture $\frac{1}{2}$ BVN $[(0,0), \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}]$ + $\frac{1}{2}$ BVN $[(3,3), \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}]$. There are two widely separated spherical clusters in this data set. The density estimates ($f_N(x) \propto n_i^2 \, WSS_i^{-1}$) over the k=40 clusters obtained by k-means, and the resulting hybrid clusters are given respectively in Figures C and D. As do most other hierarchical clustering algorithms, the hybrid method identifies correctly the two distinct modes in the population.

3. Experiment Three: In this experiment, the sample of size 1000 from the mixture ½BVN $[(0,0), \begin{pmatrix} 9 & 0 \\ 0 & 4 \end{pmatrix}]$ + ½BVN $[(0,6), \begin{pmatrix} 9 & 0 \\ 0 & 4 \end{pmatrix}]$ resembles two elliptical clusters with a moderate amount of noise points between them. The hybrid algorithm identifies the two clusters correctly (see Figures E and F), while all of the standard joining techniques like single linkage and complete linkage fail to do so.

The CPU time consumed on the IBM 370/58 in the three examples are 10.9, 12.6, 16.8, seconds respectively. Hence, the hybrid algorithm can be considered as a practicable and consistent method for identifying high density clusters.

FIGURE A : Density Estimates over the intervals between the 40 cluster means obtained by K-means. (1000 observations from ½N (0,1) + ½N(3.1).)

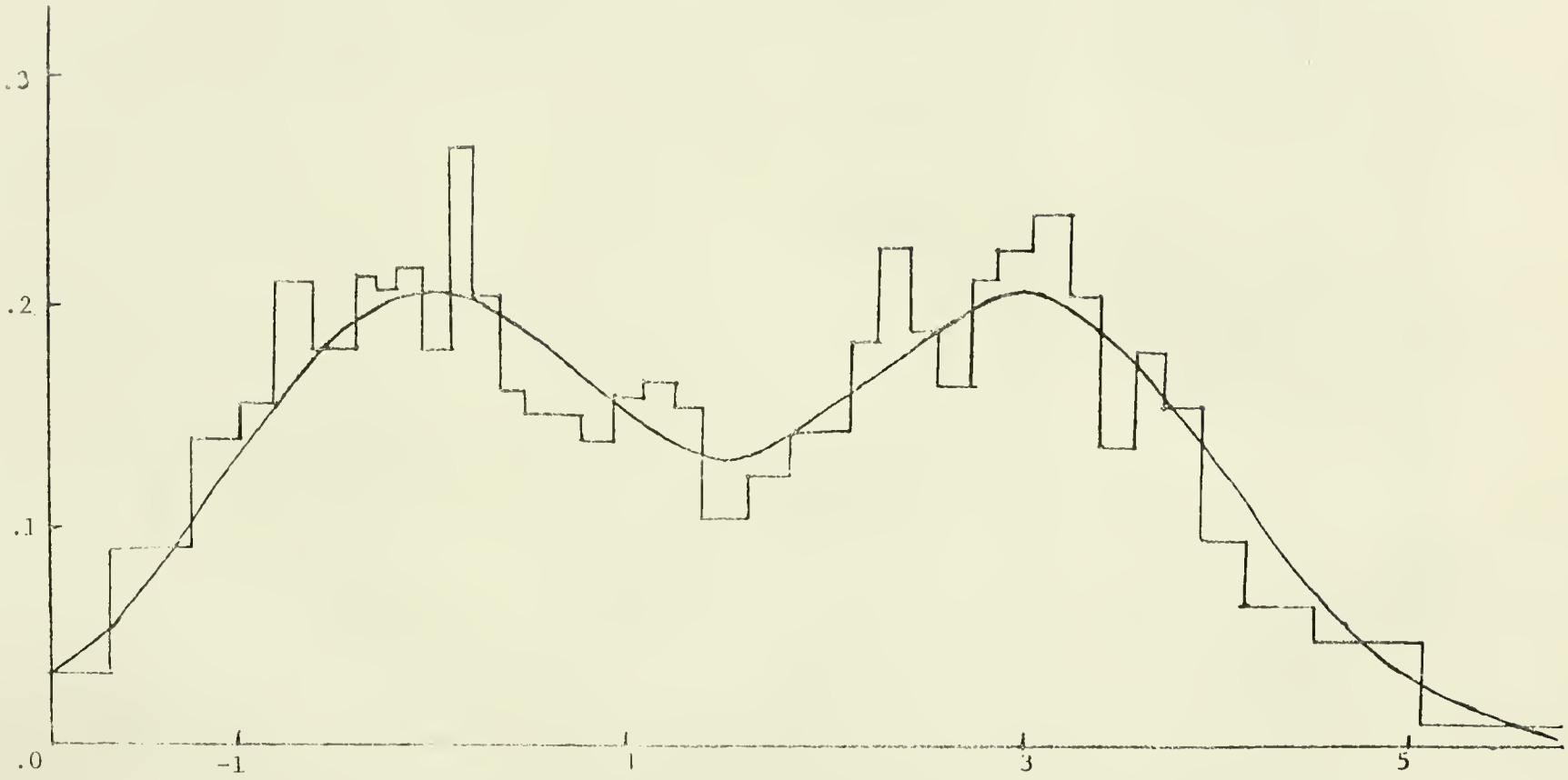


FIGURE B: Hybrid Clustering of 1000 observations from ½N (0,1) + ½N (3,1).

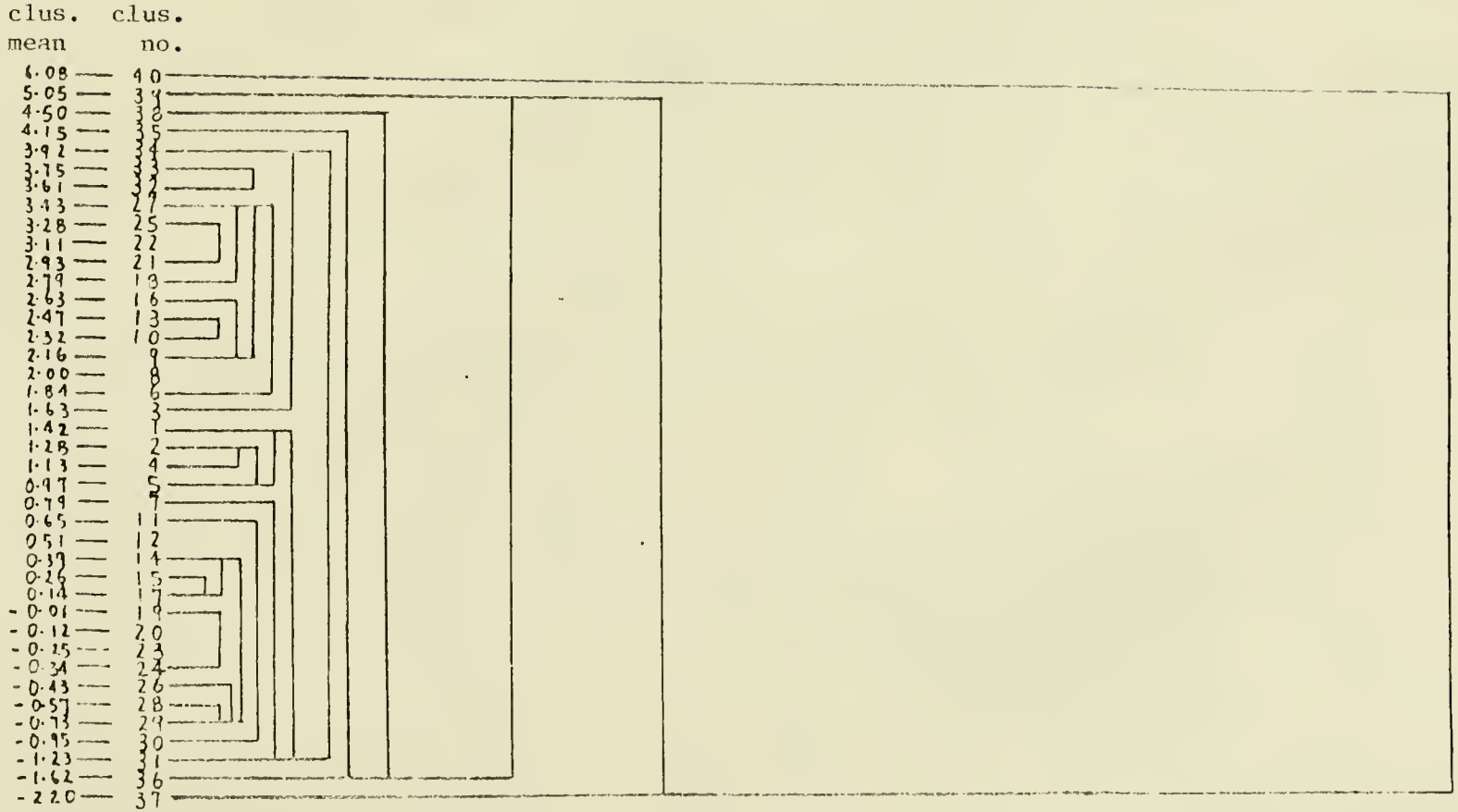


FIGURE C: Density Estimates obtained in the K-means step (k=40).

(1000 observations from $\frac{1}{2}BVN\left[(0,0), \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right] + \frac{1}{2}BVN\left[(3,3), \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right]$.)

Cluster numbers are plotted at the cluster means.

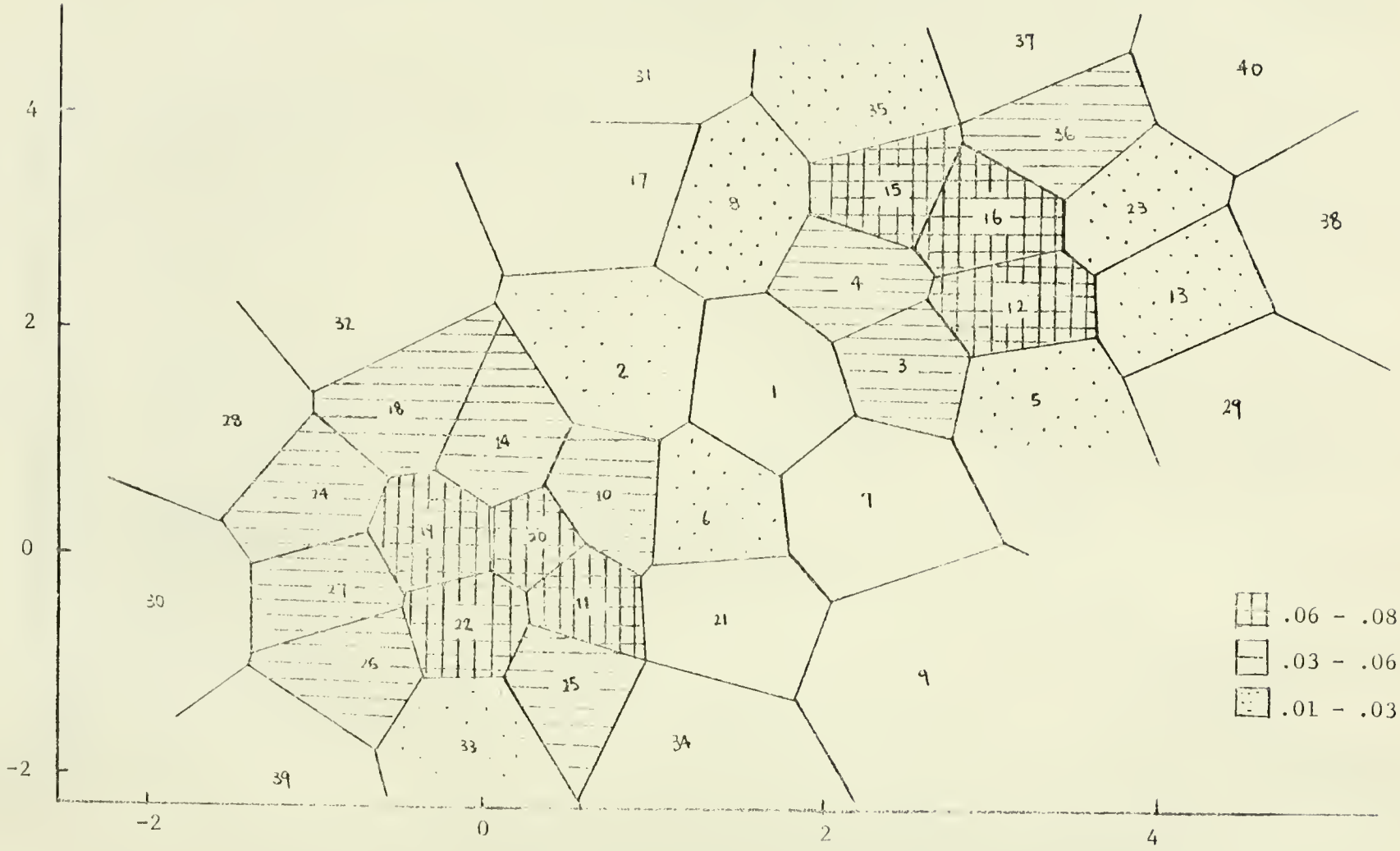


FIGURE D: Hybrid Clustering of 1000 observations from

$$\tfrac{1}{2}\text{BVN}\left[(0,0), \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right] + \tfrac{1}{2}\text{BVN}\left[(3,3), \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right] .$$

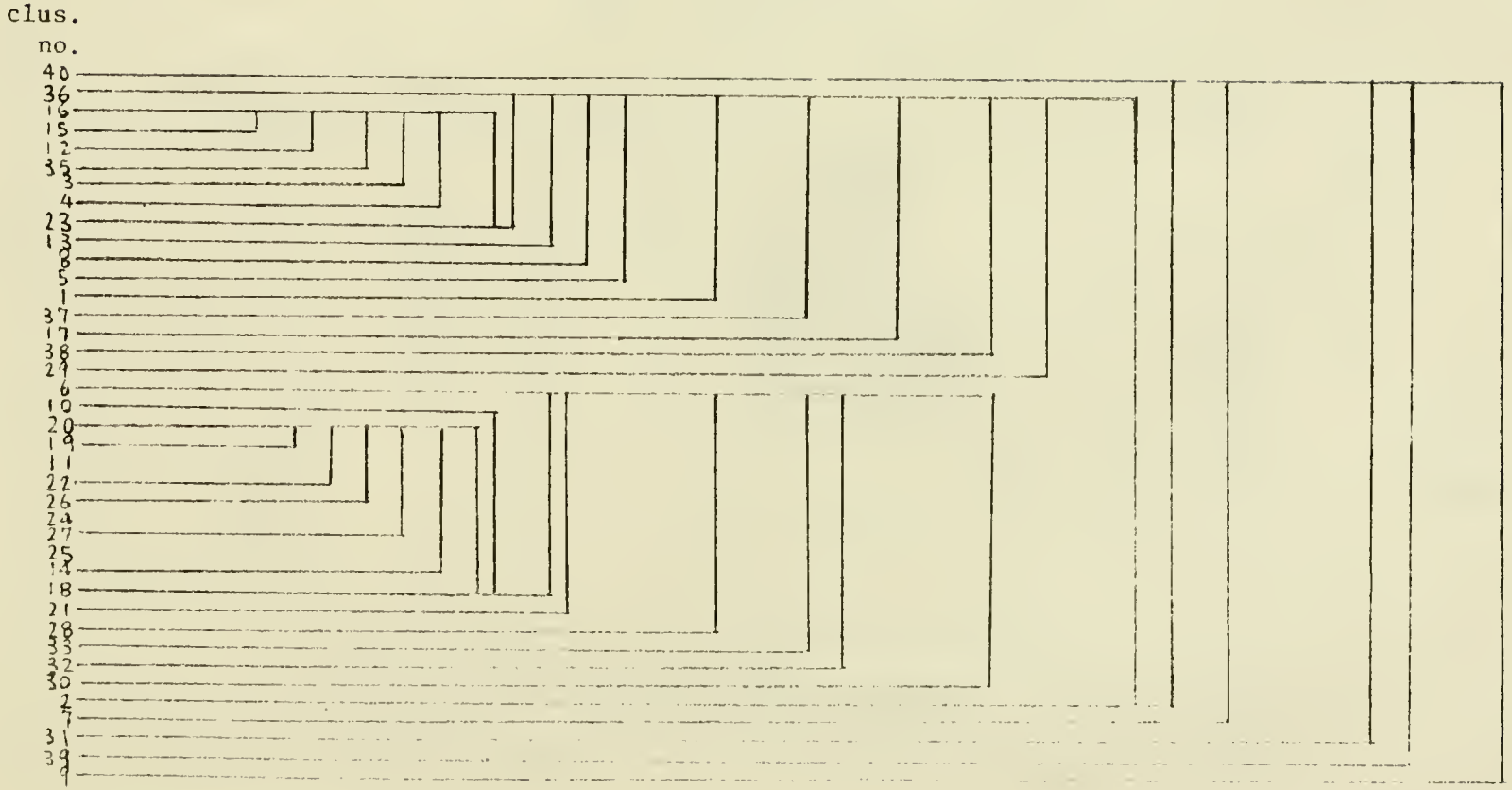


FIGURE E: Density Estimates obtained in the k-means step (k=40).

(1000 observations from $\frac{1}{2}$BVN $[(0,0), \begin{pmatrix} 9 & 0 \\ 0 & 4 \end{pmatrix}] + \frac{1}{2}$ BVN $[(0,6), \begin{pmatrix} 9 & 0 \\ 0 & 4 \end{pmatrix}]$.)

Cluster numbers are plotted at the cluster means.

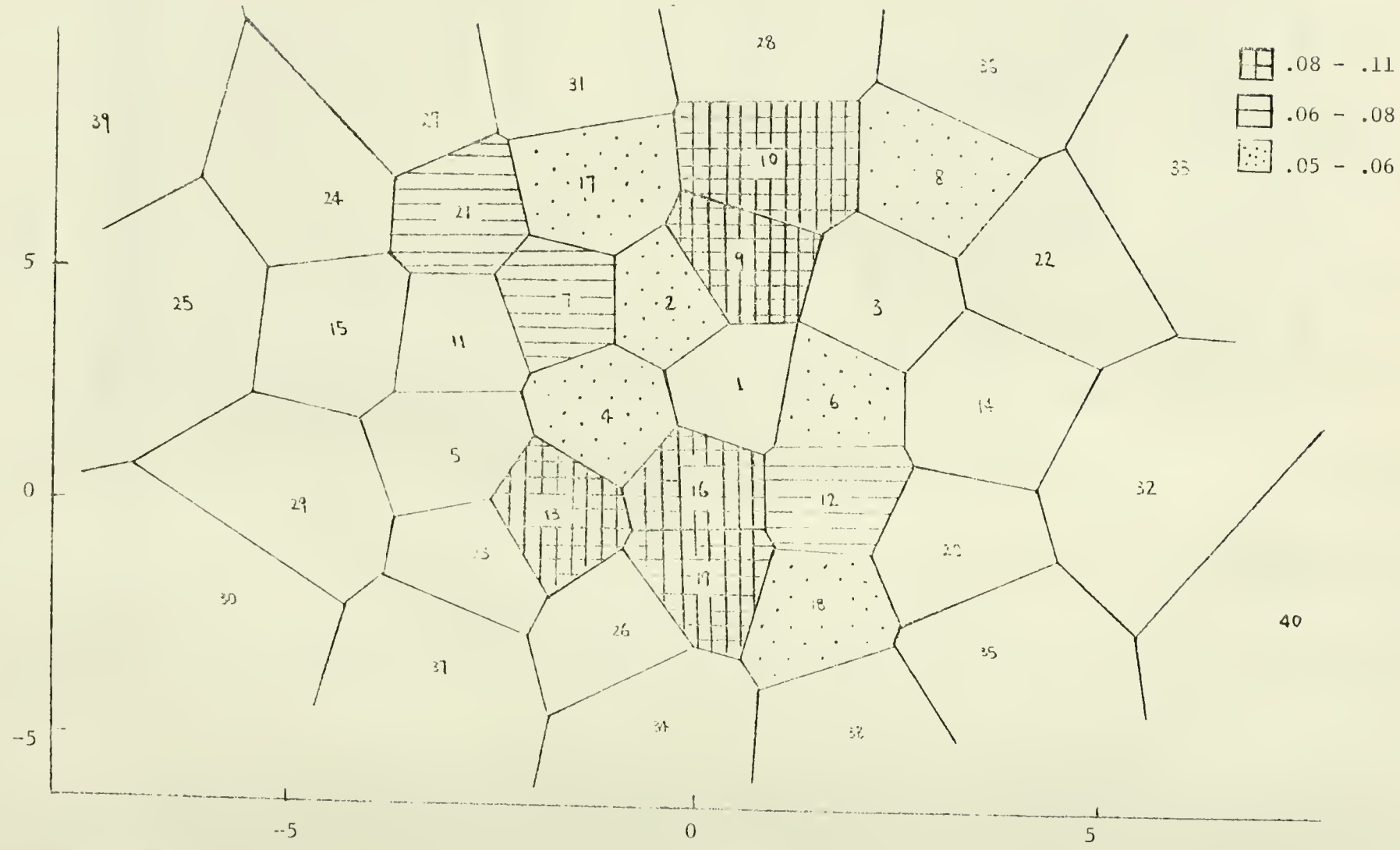
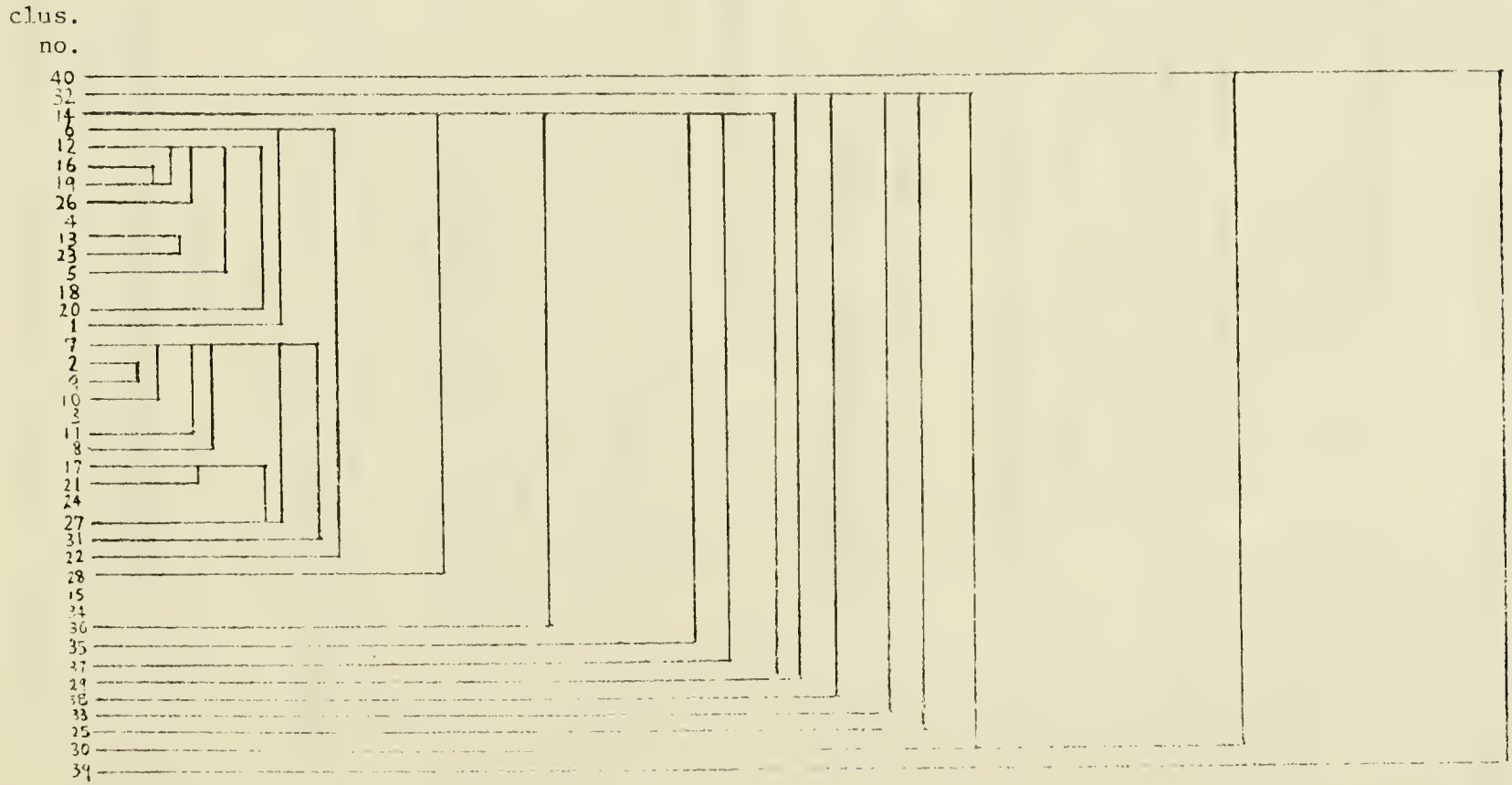


FIGURE F: Hybrid clustering of 1000 observations from
$\frac{1}{2}$BVN $[(0,0), \begin{pmatrix} 9 & 0 \\ 0 & 4 \end{pmatrix}] + \frac{1}{2}$BVN $[(0,6), \begin{pmatrix} 9 & 0 \\ 0 & 4 \end{pmatrix}]$.



References

Hartigan, J.A. (1975), Clustering Algorithms, New York: John Wiley & Sons .

_____ (1977a), "Distribution problems in clustering", in Classification and Clustering, ed. J. Van Ryzin, New York: Academic Press.

_____ (1977b), "Clusters as modes", in First International Symposium on Data Analysis and Informatics, Vol.2, IRIA, Versailles.

_____ (1978), "Asumptotic distributions for clustering criteria", Annals of Statistics, 6, 117-131.

_____ (1979), "Consistency of single linkage for high density clusters", unpublished manuscript, Department of Statistics, Yale University.

_____, and Wong, M.A. (1979), "Algorithm AS136: A K-means clustering algorithm", Applied Statistics, 28, 100-108.

Jardine, N. and Sibson, R. (1971), Mathematical Taxonomy, New York: John Wiley & Sons.

Ling, R.F. (1973), "A probability theory of cluster analysis", Journal of the American Statistical Association, 68, 159-164.

MacQueen, J.B. (1967), "Some methods for classification and analysis of multivariate observations", in the Proceedings of the Fifth Berkeley Symposium, 1, 281-297.

Pollard, D. (1979), "Strong Consistency of k-means clustering", unpublished manuscript, Department of Statistics, Yale University.

Sneath, P.H.A. (1957), "The application of computers to taxonomy", Journal of General Microbiology, 17, 201-226.

_____ and Sokal, R.R. (1973), Numerical Taxonomy, San Francisco: W.H. Freeman.

Sorensen, T. (1948), "A method of estimating groups of equal amplitude in plant sociology based on similarity of species content", Biologiske Skrifter, 5, 1-34.

Wegman, E.J. (1972), "Nonparametric probability density estimation: I. a summary of available methods", Technometrics, 14, 533-546.

Wishart, D. (1969), "Mode Analysis", in Numerical Taxonomy, ed. A.J.Cole, New York: Academic Press, 282-308.

Wong, M.A. (1979), "Hybrid Clustering", unpublished Ph.D. dissertation, Department of Statistics, Yale University.

_____ (1980), "Asymptotic Properties of the k-means algorithm as a density estimation procedure", Working Paper #2000-80, Sloan School of Management, Massachusetts Institute of Technology.